Finnish Social and Health Data Permit Authority 1 February 2021

Issued on	1 February 2021
Effective date	1 February 2021
In effect	Until further notice
Legal basis	Section 12(2) of the Act on the Secondary Use of Health and Social Data (552/2019) (referred to as the 'Secondary Use Act' in the regulation)

Regulation on data contents, concepts and data structures of data descriptions

issued by the Finnish Social and Health Data Permit Authority

1 Background

Under the Secondary Use Act, the controllers defined in the act must prepare data descriptions of the data contents of their information resources so that the suitability of the data for the uses referred to in the act can be assessed on their basis.

These controllers are listed in section 6 of the act. Public providers of health and social services operating in Finland are the largest group of actors to which the requirement applies. The Health and Social Data Permit Authority Findata (hereafter the 'Data Permit Authority') must also prepare data descriptions if it produces pre-processed data referred to in the act for its own use.

The Data Permit Authority provided the controllers in question with information on the data description work in spring 2020 and drafted this regulation in extensive consultation with controllers.

The Data Permit Authority issues this regulation under section 12(2) of the Secondary Use Act. Before entering into force, the regulation was circulated for comments and the feedback received during the process was considered when the regulation was finalised.

For the content of the regulation (data content, concepts and data structures of the data descriptions), see Appendix 1 to this regulation.

The regulation has been issued in conjunction with the Ministry of Social Affairs and Health Decree on the Obligation of the Controller and Data Permit Authority to prepare a Data Description.

2 Scope of the regulation

This regulation is the regulation of the Data Permit Authority on the data contents, concepts and data structures of the data descriptions referred to in section 12(2) of the Secondary Use Act. All controllers referred to in section 6 of the act must comply with the regulation.

3 Entry into force

This regulation will enter into force on 1 February 2021 and it will remain in effect until further notice.

Provisions on the timing of an obligation referred to in the regulation and in the act are given by Ministry of Social Affairs and Health decree.

4 Applicable legislation

Section 12(2) of the Act on the Secondary Use of Health and Social Data (552/2019).

Johanna Seppänen	Peija Haaramo
Director	Chief Specialist

This regulation has been signed electronically.

Appendix 1: Data descriptions of secondary use of health and social data

Contents

1 General information 2
1.1 Concepts 2
Data 4
Metadata 4
Data controller 4
Information resource 4
Data resource 4
Dataset 5
Variable 5
Codes and code lists 5
Glossaries 5
Application programming interface 6
Machine-readable format 6
1.2 Key principles of data description work 6
2 Tools 7
2.1 Aineistoeditori 7
2.2 Aineistokatalogi 7
2.3 Code lists 8
3 Minimum requirements for data descriptions 8
3.1 Description of data resources and datasets 8

1 General information

This document specifies the regulation on data descriptions issued in conjunction with the Act on the Secondary Use of Health and Social Data (552/2019). Under section 12 of the act, the organisations listed in section 6 of the same act must, in their capacity as controllers, prepare data descriptions of the data contents of their information resources to the extent that these contents fall within the scope of the Secondary Use Act. This ensures that the suitability of the register data in question for the uses listed in section 2 of the Secondary Use Act can be assessed. These uses are as follows: statistics, scientific research, development and innovation activities, education, knowledge management, steering and supervision of social and health care by authorities, and planning and reporting duty of an authority.

The Health and Social Data Permit Authority Findata prepares data descriptions of the pre-processed data in its possession.

The purpose of the regulation on data descriptions is

1. to ensure uniform and high-quality description of the data resources of the organisations referred to in section 6 of the Secondary Use Act
2. to allow safe and effective use of health and social data resources and to ensure that the Data Permit Authority can perform its tasks and provide its customers with services in an effective and high-quality manner in accordance with the principles of good governance, and
3. to promote the interoperability of the information resources of different organisations.

This appendix contains more detailed instructions on the data contents, concepts and data structures of the data descriptions. Before issuing the regulation, the Data Permit Authority consulted relevant organisations within the framework of an open and public consultation round in November and December 2020. The Act on Information Management in Public Administration (Act on Information Management 906/2019) has been taken into account in the drafting of this regulation. The regulation will be periodically updated.

The data description instructions contain references to the following acts:

- Language Act (423/2003)
- Administrative Procedure Act (434/2003)
- Act on the Secondary Use of Health and Social Data (552/2019)
- Act on Information Management in Public Administration (906/2019)

1.1 Concepts

The data contents of the registers falling within the scope of the Secondary Use Act are described in the data description work. In this regulation, data descriptions are considered to have three main levels: data resource, dataset and variable (see Figures 1 and 2). The key concepts used in data description work are defined below. In the definition, they are given the meanings that are used for them in this appendix. The Act on Information Management and the thesaurus and ontology service Finto have been the main sources used in the definition work.

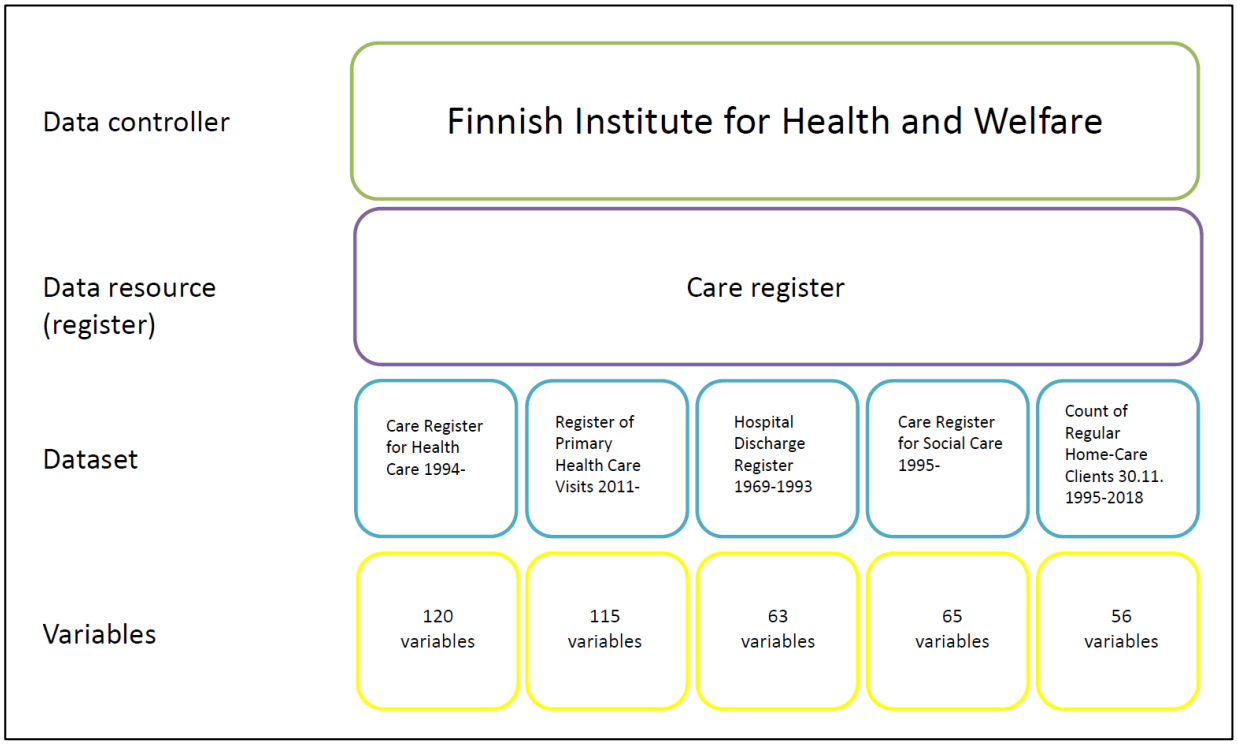

Figure 1. Example of data description levels: care register of the Finnish Institute for Health and Welfare (THL)

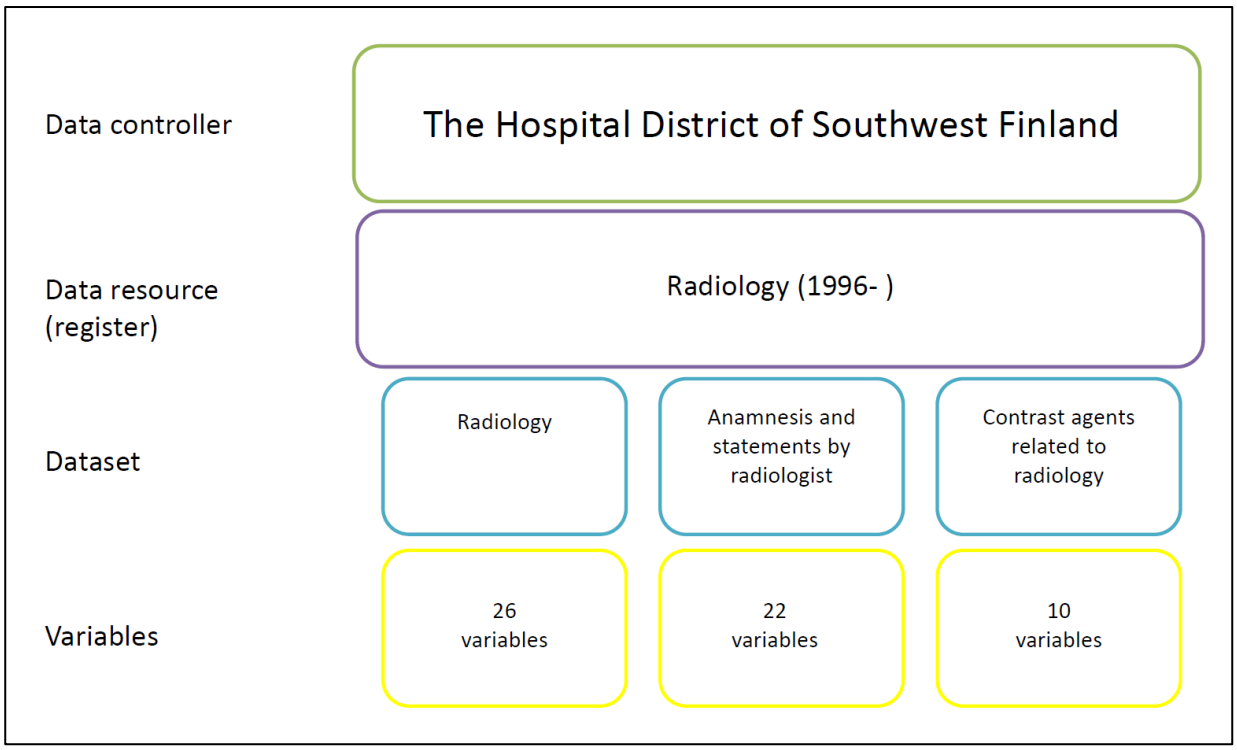

Figure 2. Example of data description levels: imaging examinations carried out in the Southwest Finland Hospital District

Data

Data is information at the lowest processing level. Even though it may not be possible to interpret data, it can nevertheless serve as a source of information when processed.

- Data of many different types falls within the scope of the Secondary Use Act and the aim is to describe all this data in the most appropriate manner.

Metadata

Metadata contains information about data, and it describes the context, content or structure of data resources.

Data controller

Controller means the authorities and organisations listed in section 6 of the Secondary Use Act.

Information resource

An information resource is a collection of data resources used by the controller in their tasks or other activities that is processed manually or by means of information systems.

- A data resource or a collection of data resources compiled for a specific purpose and comprising data that is logically or physically interconnected.

Data resource

A data resource is a collection of data used by the controller to perform a specific task or to provide a specific service.

- A data resource is also an identifiable entity consisting of data stored in a data medium.
- A data resource forms part of the information resource used by the controller. In some cases, however, an information resource may only comprise a single data resource.
- A data resource may consist of several datasets.
- A data resource is a set of data containing information and describing a specific phenomenon or specific events. It is a logical entity based on a specific purpose.
- For example, a statutory register may constitute a data resource. The purpose of the register is to serve as a systematic resource containing the data on the registered subjects required for a specific purpose.

Dataset

A dataset is a method for grouping data within a data resource on the basis of specific principles.

- A data resource is a thematic entity within which the parts essential for the user are stored as a single entity or as datasets.
- Discretion must always be used when a data resource is divided into datasets and the criteria for dividing the resource must also be applied on a case-by-case basis.
- When a data resource is divided into datasets, the requirements of the parties managing the resource and, in particular, its users should be considered: how the users perceive the structure of the data resource and which approach is best for them.
- It is not necessary to structure all data resources used by the controller in a uniform manner.

Variable

Data resources and their datasets consists of variables. The data content of the data resources is stored in the variables.

- A variable refers to a feature of the object examined in the process and it varies from one unit or measurement to another.
- Variables can be numerical or non-numerical, while numerical variables can be continuous or discontinuous (discrete variables).
- Codes of the discrete variables must also be given.

Codes and code lists

A code means the value that can be assigned to a discrete variable in a data resource and the description of this value. A code set is a collection of codes with a specific purpose.

- Value or identifier, name, description and reference code set are the key characteristics of a code.
- A code list is a set of data comprising specific codes and the description data associated with the code set itself.

Glossaries

A glossary refers to terminology on a specific topic.

- As a rule, free-form texts based on the controllers' needs and practices can be used in the descriptions.

- However, it is recommended that standardised index term lists (such as YSO, TERO and FinMesh) are used in the description work.

Application programming interface

A technical interface (API) is a data transfer solution allowing electronic exchange of information between two or more information systems.

Machine-readable format

Machine-readable format is a file format structured so that computer programs can easily identify data resources and pieces of data and their structures and extract them.

1.2 Key principles of data description work

The main aim of the work is to produce a comprehensive national catalogue of the descriptions of health and social data resources. The catalogue is intended for persons and parties using these data resources for the purposes specified in the Secondary Use Act.

The controller must describe all its data resources falling within the scope of the Secondary Use Act, from data resource level down to variables and code sets. The work should be performed in stages so that each controller starts the description work from the data resources that are in particularly high demand among secondary data users. The controller should turn to the Data Permit Authority for advice if it is not clear which data resources fall within the scope of the Secondary Use Act or where the description work should be started.

The Data Permit Authority cooperates with controllers but does not produce data descriptions for them. The Data Permit Authority provides controllers with advice and training in the production of data descriptions and disseminates information on matters concerning data descriptions.

The data descriptions must be prepared and stored in electronic format. As a rule, an organisation must store all its data descriptions in the same place. All data descriptions must be made publicly accessible as soon as they are finalised (excluding secret metadata). The aim is to promote the use of persistent identifiers in data descriptions.

When producing data descriptions, the controller must take into account all applicable provisions of the Language Act and comply with them. With the metadata / data description editor Aineistoeditori, descriptions can be prepared in both Finnish and Swedish and the tool can also be used to produce data descriptions in English, should the controller wish to do that. Regardless of the language choice, the expressions used in the data descriptions must be clear and easy to understand for secondary data users.

2 Tools

This section contains information on the tools that are recommended to be used in the description work.

The data description editor Aineistoeditori, data catalogue Aineistokatalogi and the catalogue view of the Data Permit Authority are the primary tools to be used in the description work.

2.1 Aineistoeditori

Findata recommends using Aineistoeditori to describe data resources, datasets and variables falling within the scope of the Secondary Use Act. Aineistoeditori is a Finnish-developed data description system specifically intended for data description work for the purposes of the Secondary Use Act. Aineistoeditori is managed by the Data Permit Authority and the Finnish Institute for Health and Welfare and it can be used free of charge.

Aineistoeditori is an open-source service, and its code is available in GitHub. The applicable parts of its metadata model are compatible with the main international metadata standards of the sector.

Aineistoeditori only contains descriptions / metadata of data, not actual data). The details of data resources and datasets must be fed in Aineistoeditori manually though variables can and should be entered in the system as CSV files. Work is also under way to further develop the option of interface import.

Aineistoeditori also contains a feature allowing the production of the records of processing activities required under the General Data Protection Regulation of the EU. Data controllers can use this feature if they wish to do so.

The Data Permit Authority provides controllers with advice and instructions on the use of Aineistoeditori.

- Location of Aineistoeditori: https://aineistoeditori.fi/(the website is in Finnish)
- Instructions for use: https://yhteistyotilat.fi/wiki08/x/My6wAg (the website is in Finnish)
- Source code: https://github.com/THLfi/thldtkk
- Login: Employees at Virtu agencies should log in using Virtu, while other users can log in using the THL authentication service. To get IDs and passwords for the service, contact info@aineistokatalogi.fi.

2.2 Aineistokatalogi

All data descriptions must be made accessible in the public information network as soon as they are finalised. It is easy to publish the descriptions prepared with Aineistoeditori in Aineistokatalogi. Aineistokatalogi is managed by the Data Permit Authority and the Finnish Institute for Health and Welfare. Aineistoeditori and Aineistokatalogi also contain descriptions of data resources outside the

scope of the Secondary Use Act and for this reason, the Data Permit Authority will create a separate catalogue view displaying only the data resources falling within the scope of the Secondary Use Act.

Aineistokatalogi has been developed as a national publishing platform for data descriptions and it can be used free of charge. The descriptions published in Aineistokatalogi are available to other applications through an open interface. Aineistokatalogi contains descriptions at data resource, dataset and variable level as well as the search function for variables.

- Location of Aineistokatalogi: https://aineistokatalogi.fi (the website is in Finnish)

2.3 Code lists

A variety of different tools can be used to describe code sets and the main principles for describing them are as follows:

- The codes/classification used by each discrete variable must be given.
- Codes published on open platforms that can be provided with a hyperlink must be used.
- The information on the code lists should primarily be saved in machine-readable format.
- Existing code lists can and should be used if they meet these requirements.

- Suitable tools include the following:
- The reference data tool maintained by the Digital and Population Data Services Agency provides a technical platform for the use and maintenance of joint public administration code lists and classifications: https://koodistot.suomi.fi and https://wiki.dvv.fi/pages/viewpage.action?pageId=21779546 (websites are in Finnish)
- The code server maintained by Kela and the Finnish Institute for Health and Welfare: https://koodistopalvelu.kanta.fi

3 Minimum requirements for data descriptions

The information listed in this document must be included in the descriptions of data resources falling within the scope of the Secondary Use Act. The descriptions must be at the level of data resources, datasets, variables and (if necessary) code lists. Aineistoeditori also has other data fields that controllers can use. Priority should be given to data resources in use and regularly updated data resources, but the controller may also include past data in the description.

3.1 Description of data resources and datasets

The data on data resources and datasets given in the table below must always be included in the data descriptions. If the data resource is not divided into datasets, the details of the data resource

must also be included in the description of the dataset. If needed, the Data Permit Authority may provide the controller with advice on how to divide data resources into datasets.

In Aineistoeditori, data resources belonging to the same entity can be combined into a series. The controller can freely complete most of the data fields but in a few places, ready-made menus with glossary-based options are provided instead.

Data field	**Description**
Name	Data resource name (informative and unique) The year must also be given if for example, it is a question of an annual extraction of information contained from a larger entity.
Description	Concise summary of the data resource. Recommended maximum length 1,500 characters. A concise description of the data resource contents and why, how, where and when the data has been collected and from which sources. The subsequent fields in which questions on the data resource are asked should be considered so that repetition can be avoided.
Organisation	The organisation responsible for the data resource at the time of the description. If the data is stored in a personal data file, the controller must be given as the organisation. Other organisations involved in the production of the data can be listed in the data description field. *This information may not be entered separately for datasets.*
People / contact details related to the data resource	At least the contact person/other contact details must be given (for example, the public/shared email address of the controller's advisory service). Persons can also be named for the following roles: 1. Sample technician (responsible for the samples included in the data resource) 2. Register manager (responsible for the register used as the data resource/included in the data resource) 3. Author (author of the data resource) 4. Processor (person processing the information contained in the data resource) 5. Data manager (responsible for the accuracy of the data and the extractions from it) 6. Statistical manager (responsible for the statistics produced from the data resource) 7. Study director (in charge of the study in which the data was collected) 8. Study coordinator (responsible for the study in which the data was collected) 9. Contact person (responsible for data-related inquiries) A person with more than one role must be registered separately for each of the roles.

Data field	Description
Links	Details of the data-related websites and links relevant to the user (such as the link to the organisation's advisory service and key data-related publications). The link text and its URL must be given separately. *Example 1:* Link text: For more information, visit thl.fi URL: https://www.thl.fi/en_US/tilastot/tiedonkeruut/hoitoilmoitusjarjestelma-hilmo *Example 2:* Link text: Specialised psychiatric care 2019 statistical report URL: http://urn.fi/URN:NBN:fi-fe20201217101056
Terms of use	Determine how the data resources are made available to parties outside the organisation. There are five options to choose from and, as a rule, the data resources falling within the scope of the Secondary Use Act belong to class 2 (data permit): 1. Open data: the data resources are openly accessible 2. Data permit: a data permit is required to access the data resources 3. Biobank data permit: the data can be made available by separate agreement or consent 4. Agreement: the data can be made available for research or other use under a cooperation agreement; applications are considered on a case-by-case basis 5. No disclosure: no data is disclosed to outsiders
Target population	Population covered by the data. The information must be as specific as possible: description of the observational units, any geographic and temporal demarcation, and any inclusion and exclusion criteria. *Examples:* 1. Clients of inpatient care and day surgery in public and private health care and clients of public specialised outpatient care. Clients of public primary health care (health centres) and clients of institutional care and housing services on social care. 2. Live births and stillbirths of foetuses with a birth weight of at least 500 g or with a gestational age of at least 22 weeks. Mothers of the above-mentioned children. 3. Patients in inpatient care and periods of care in Finnish health care in 2015 (excluding residents of Åland). 4. Recipients of social assistance in Southwest Finland in 2014.
Regional coverage	Determines the geographic coverage of the data with maximum accuracy. *Examples:* 1. Helsinki 2. Kainuu region 3. Finland

Data field	Description
Reference period start date	Reference period is the period covered by the information contained in the data resources. Note that the reference period does not indicate when the data has been coded or the documents converted into machine-language format.
Reference period end date	Reference period is the period covered by the information contained in the data resources. If the data collection process continues, no end date is required.
Data resource lifecycle phase	Options: 1. Analysis and reporting stage: data resources are in active use 2. Archived: data resources have been archived 3. Destroyed: data resources have been destroyed 4. Planning stage: data resources are in the planning stage; data collection has not yet started 5. Data collection stage: data collection is under way
Keywords	The keywords describing the main contents and form of the data resources from Finto ontologies (health and welfare ontology TERO, YSO and FinMesh are recommended).
Additional keywords	Give at least “toisiolaki” as keyword in this field. This acts as an identifier indicating that the data resource falls within the scope of the Secondary Use Act. You can also add other keywords describing the key content and form of the data resource that cannot be found in ontologies.
Type of data resource	Select the type/types that best describe the data resource. Options: 1. Customer record data 2. Biobank data 3. Interview data 4. Observational data 5. Survey data 6. Other documents 7. Sample/specimen data 8. Patient record data 9. Registry data 10. Statistics data 11. Population surveys